

Stochastic Shortest Path with Energy Constraints in POMDPs

Tomáš Brázdil
Masaryk University
Brno, Czech Republic
xbrazdil@fi.muni.cz

Krishnendu Chatterjee
IST Austria
Klosterneuburg, Austria
kchatterjee@ist.ac.at

Martin Chmelík
IST Austria
Klosterneuburg, Austria
mchmelik@ist.ac.at

Anchit Gupta
IIT Bombay
Mumbai, India
anchit@iitb.ac.in

Petr Novotný
IST Austria
Klosterneuburg, Austria
pnovotny@ist.ac.at

Abstract

We consider partially observable Markov decision processes (POMDPs) with a set of target states and positive integer costs associated with every transition. The traditional optimization objective (stochastic shortest path) asks to minimize the expected total cost until the target set is reached. We extend the traditional framework of POMDPs to model energy consumption, which represents a hard constraint. There are energy levels that may increase and decrease with transitions, and the hard constraint requires that the energy level must remain positive in all steps till the target is reached. Our contribution is twofold. First, we present a novel algorithm for solving POMDPs with energy levels, developing on existing POMDP solvers and using real-time dynamic programming as its main method. Our second contribution is related to policy representation. For larger POMDP instances the policies computed by existing solvers are too large to be understandable. We present an automated procedure based on machine learning techniques that automatically extracts important decisions of a policy and computes its succinct, human readable representation. Finally, we show experimentally that our algorithm performs well and computes succinct policies on a number of POMDP instances from the literature that were naturally enhanced with energy levels.

1 Introduction

Motion and task planning for autonomous agents are one of the classical problems studied in AI and robotics. One of the main challenges that make the problem difficult is the presence of *uncertainty* about the state of the agent and its environment [33], caused for instance by the agent’s unreliable sensors. To account for these issues, powerful abstract frameworks for solving

planning problems under uncertainty were developed, among which the *Partially observable Markov decision processes (POMDPs)* play a crucial role.

Each POMDP describes a discrete, typically finite-state system that exhibits both probabilistic and non-deterministic behaviour [54, 41]. Probabilities are useful for modelling sensor errors, hardware failures, and similar events whose rate of occurrence can be established empirically, while non-determinism represents the freedom of the agent’s controller to choose appropriate control input. The imperfection of agent’s sensors is represented by *observations*. In every step the controller receives an observation but not the current state itself. *Policies (policies)*, i.e. rules for resolving non-determinism in POMDPs, can be viewed as blueprints for implementing concrete controllers of the agent. Hence, given a POMDP modelling an agent and its environment the usual task is to find a policy ensuring that the behaviour of the system conforms to a given *specification* or *objective*.

Various types of objectives in POMDPs have been studied. Typically it is assumed that there is a reward (resp. cost) function that assigns rewards (resp. costs) to transitions of the system. The goal of the agent is to maximize (resp. minimize) the reward (resp. cost) over a finite-horizon [52], or over an infinite horizon [55], where the sequence of rewards (resp. costs) can be aggregated by considering the *discounted reward* [42, 53] or the *average reward* [43, 21], etc. Particularly relevant from the planning point of view is the *indefinite-horizon (or stochastic shortest path)* objective [8, 5, 17], which asks to compute a policy that reaches a state from a given set of target states T and minimizes the expected total cost till the target set T is reached, i.e., the expected sum of costs of all transitions traversed before reaching T . Typically T is such that reaching a state of T corresponds to the agent completing some assigned task.

Energy and Soft vs. Hard Constraints Most autonomous robotic devices operate under certain *energy constraints*, i.e. they need a steady supply of some resource (in the form of, e.g. fuel, electricity, etc.) to operate correctly. While the stochastic shortest path (SSP) objective can in principle express specifications of the form “complete the task while minimizing the expected consumption of some resource,” this approach is not suitable for modelling of resource-constrained systems, as the SSP objective only talks about the expected cost, without giving any guarantees on the cost of concrete executions of the modelled system, i.e., it is an example of a *soft constraint*. In particular, to use costs in SSP objectives to model resource consumption we would need to assume that the amount of a resource consumed by making a transition (represented by the transition’s cost) is always available. This is not always realistic. An autonomous robot typically has a battery of a finite capacity (or a fuel tank of finite volume) which is continually depleted as the robot operates. The total amount of a resource required to complete the task can exceed this capacity, prompting the robot to periodically recharge the battery (or refuel the tank) at special charging points (petrol stations). When the resource is depleted, no action remains available to the robot, i.e., keeping the energy level positive is a *hard constraint* that must hold along every single execution of the system, no matter the outcomes of stochastic choices. (The issue of expectation-based vs. execution-based constraints was already examined in the setting of perfectly observable MDPs, see [1, 50, 51, 49].)

In this paper we address this issue and extend POMDPs with *energy constraints*. That is, to a POMDP \mathcal{M} with a given objective we assign a positive integer capacity *cap* and to each transition of \mathcal{M} we assign an integer *update* representing the amount of a resource consumed or reloaded by this transition. Such a POMDP starts with some initial level of a resource (say *cap*, i.e. the resource is loaded to full capacity) which is then modified as the system evolves: whenever a transition with some update *u* is traversed, the resource level changes from ℓ to $\min\{\ell + u, \text{cap}\}$ (discarding any quantity exceeding *cap* captures the fact that the robot’s storage capacity cannot be exceeded). The task is to find a policy ensuring the original objective and at the same time ensuring that the resource level stays positive till the target is reached.

Our Results on POMDPs with Energy Constraints We study *energy-reachability* problem for POMDPs. In the qualitative version of the problem we ask to find a policy that ensures that the expected total cost is finite before reaching the target state and at the same time keeps the resource level positive. In the quantitative version we additionally seek for a pol-

icy that, on top of above two conditions, minimizes the expected total cost till the target is reached. We show how to solve both these problems by reducing them to corresponding problems in POMDPs without energy constraints. In particular, we show that the qualitative energy-reachability problem is EXPTIME-complete, i.e. it has the same complexity as unconstrained qualitative reachability. We experimentally evaluate our approach on standard POMDP models of robot planning.

Representation of Policies Solving POMDPs with energy-reachability objectives highlights another relevant issue: the representation the computed policies. Policies in POMDPs are often represented in a form of a table [8] or *plan graphs* [33], which are equivalent to so called finite-memory policies used in verification [20]. Size of these structures can become very large and not very readable by humans. For instance, policies ensuring that a target state is reached with probability 1 might require table or plan graph of size exponential in the size of the POMDP [19].

There are two reasons why size and representation of policies matter. First, as offline-computed policies have to be implemented on real-world devices, it is advisable to reduce their memory requirements so that they fit into the device’s memory and do not cause delays through inefficient memory access. The second issue, to which we devote a particular attention in this paper, is the one of human readability. From the engineering point of view it is vital to be able to visualize the policy and understand its behaviour. This is reflected in numerous informal rules for safety-critical system design that enforce “simplicity” and “readability” [29, 57] as well as in academic treatments of the subject [35, Chapter 2 on “Simplicity”]. Although many methods for policy computation in POMDPs produce results that are correct by design, the behaviour induced by the computed policy in an actual device might not be reasonable, due to either using an improper model of the system, or too weak specification that does not rule out all undesirable behaviours. In such a case, comprehension of the computed policy can lead the system designer to refine the model or specification in an appropriate way. Easily understandable descriptions of policies can be also interesting for type approval authorities.

Readability of policies is relevant for POMDPs in general, but the issue is especially pronounced in energy-constrained POMDPs, as the standard representation does not reveal which decisions depend on states and which depend on current resource level, an information useful for identifying bottlenecks caused by insufficient storage capacity or exploiting the fact that policy’s dependency on resource levels might not be complex (e.g. “when low on fuel, go to a gas station”).

Our Results on Policy Representation We study succinct representation of policies via *decision trees*.¹ A decision tree (DT) is an easily visualisable data structure in which leaves represents actions prescribed by the policy and branching in internal nodes represents decisions that the policy makes in order to select a suitable action. To obtain a DT-representation from the corresponding table representation we utilize machine learning techniques for learning DTs. There advantage of this approach is that learning algorithms are often able to exploit the structure of the model, identify the crucial decisions made by the policy and encode *only these* decisions in the DT. This typically results in much more succinct representation without significant loss of the policy’s performance. To support this claim, we present experimental results on learning DT-represented policies using several well-known learning tools. As discussed in the previous paragraph, our approach can be seen as a generic technique for POMDPs which is particularly apt for use in the presence of energy constraints.

2 Related Work

Our model of POMDPs with energy constraints, where the goal is to optimize the expected total cost while ensuring constraints on resource consumption, resembles the standard framework of constrained POMDPs [30, 59, 34] a generic framework for enforcing constraints in POMDPs which has received considerable attention in various application domains [61, 31, 60] (see also [1] for related concepts in the setting of perfectly observable MDPs). The crucial difference between constrained POMDPs and our energy constraints is that the constraints in constrained POMDPs are soft, i.e. they are bounds on the *expectation* of some quantity, while we require that the resource level stays always positive (not just on average) in all runs (see also a discussion in Section 1).

As mentioned earlier, we extend the previous work on indefinite-horizon objective [8, 5, 17] by adding energy constraints. Our notion of energy constraints is similar to the one used in verification, in particular to so called *energy games and MDPs* [16, 18] and *consumption games* [11], although none of these concepts was considered in a partially observable setting so far. DTs have already been successfully used to represent policies in verification of perfectly observable MDPs modelled in the well-known PRISM tool [10]. For POMDPs, in [9] they consider a situation where the POMDP itself is encoded succinctly using DTs and similar structures, and they use this assumption to design a specific algorithm

computing a desired policy (which itself is not encoded as a DT). In contrast, we assume that the model is given explicitly and use generic machine learning methods to infer succinct representations of policies. In [7] they study relationship of DTs and POMDPs from an inverse perspective, POMDPs are used as a tool for learning decision trees from generic datasets. DTs were also used to represent policies in a reinforcement-learning setting [25], where the agent has no a priori model of the environment.

The need for succinct and efficient representation of policies motivated the study of finite-state controllers (FSCs) in POMDPs [22, 28, 37, 24]. Intuitively, the approach is based on direct search for a small policy represented as a finite (possibly stochastic) transducer whose transitions are labelled by observations. In every step, the state of the transducer changes according to the transition function and latest observation received, and the controller then outputs an action to be performed based on the current state of the transducer. While this approach was shown to produce small and well-performing policies, we argue that our approach, while having similar goal, is conceptually different and offers an orthogonal set of advantages. The main difference is that FSCs represent a function whose domain are histories of actions and observations: each state of the finite transducer implicitly carries an information about the set of histories that lead the transducer to this state. The transducer thus captures an operational aspect of a policy, i.e. the way in which it is executed as a program. On the other hand, DTs represent functions whose domain is the set of beliefs: given a belief, a single root-leaf traversal of a DT is used to establish an action to be performed. Thus, DTs capture the logic of agent’s decision in a concrete time instant; it is up to the agent to keep an (accurate or approximate) representation of its belief (which can be done using standard computations) and thus to take care of the history-dependent aspect of decision making. The latter approach more explicitly captures the decision making process as a human-like inference of suitable action from available information, and thus we believe that it provides better readability. Another advantage of our approach, which is validated by our experiments, is that the machine learning techniques we use are able to automatically identify the “most important” decisions that amount for the majority of optimization efforts. Finally, we show that FSCs can be prone to storing an unnecessary amount of data about resource levels. We further explain differences between the two formalism on a concrete example at the end of Section 5.

One crucial difference between previous approaches to policy succinctness in both POMDP [22] and other settings [25] is that in previous work they concurrently optimize both the performance of a policy and its size,

¹These should not be confused with *policy trees* that represent a complete behaviour of a POMDP under a fixed policy [33].

which requires dedicated algorithms, while we separate these tasks: first we search for a well-performing, though possibly "ugly" policy, and then learn its succinct representation (similar approach was used in [23], where policies computed by point-based methods were "compiled" into FSCs). Thus, we present a framework for obtaining succinct representations in which various state-of-the-art algorithms for POMDP solving and DT learning can be used. On the POMDP side, this allows us to keep up with advances in solving of large POMDPs. On the DT side, we can use well-developed machine learning tools that already offer a selection of methods for tree pruning and visualization, which is important for readability.

3 Preliminaries

Notation We use $\mathbb{N}_0, \mathbb{N}, \mathbb{Z}$ to denote the sets of non-negative, positive, and all integers, respectively. For $n \in \mathbb{N}$ we denote by $[n]$ the set $\{1, \dots, n\}$. Let X, Y be finite sets. For a function $f: X \rightarrow Y$ and sets $X' \subseteq X, Y' \subseteq Y$ we denote by $f(X')$ the image of X' under f , i.e. the set $\{y \in Y \mid \exists x \in X': f(x) = y\}$ and by $f^{-1}(Y')$ the pre-image of Y' under f , i.e. the set $\{x \in X \mid f(x) \in Y'\}$. We denote by $\mathcal{D}(X)$ the set of all probability distributions on X , i.e. of all functions $f: X \rightarrow [0, 1]$ s.t. $\sum_{x \in X} f(x) = 1$. For $f \in \mathcal{D}(X)$ we denote by $\text{supp}(f)$ the *support* of f , i.e. the set $\{x \in X \mid f(x) > 0\}$. A probability distribution f is *Dirac* if $|\text{supp}(f)| = 1$. An encoding size of an object O (i.e. the number of bits needed to represent O) is denoted by $\|O\|$.

POMDPs A *Partially Observable Markov Decision Process (POMDP)* is a tuple $\mathcal{M} = (S, A, \delta, Z, O, \lambda)$ where: S is a finite set of *states*; A is a finite alphabet of *actions*; $\delta: S \times A \rightarrow \mathcal{D}(S)$ is a *probabilistic transition function* assigning to every state-action pair a probability distribution over the successor states (i.e. $\delta(s, a)(s')$ denotes the probability of making a transition from s to s' under action a); Z is a finite set of *observations*; $O: S \times A \rightarrow \mathcal{D}(Z)$ is a *probabilistic observation function* assigning a probability distribution over observations to every state-action pair; and λ is an *initial probability distribution* over the states of \mathcal{M} . We write $\delta(s'|s, a)$ as a shorthand for $\delta(s, a)(s')$.

Remark 1 (Deterministic observation function). *We remark that deterministic observation functions of type $O: S \rightarrow Z$ are sufficient in POMDPs. Informally, the probabilistic aspect of the observation function is captured in the transition function, and by enlarging the state space with the product with the observations, we obtain an observation function only on states [17]. Thus in the sequel without loss of generality we will always*

consider observation function of type $O: S \rightarrow Z$ which greatly simplifies the notation.

Runs and Histories A *run* (finite or infinite) in a POMDP is an alternating sequence of states and actions $s_0, a_1, s_1, a_2, s_2, \dots$ such that $s_0 \in \text{supp}(\lambda)$ and for every $i \geq 0$ it holds $\delta(s_{i+1}|s_i, a_{i+1}) > 0$. To a run $w = s_0, a_1, s_1, a_2, \dots$ we assign an *observed run*, i.e. a corresponding observation-action sequence $O(w) = O(s_0), a_1, O(s_1), a_2, \dots$. A *history* (finite or infinite) is an alternating sequence of observations and actions denoted as $\rho = z_0, a_1, z_1, a_2, z_2, \dots$, such that there exists a run w for which $\rho = O(w)$ (note that we already assume that function O is deterministic as noted in Remark 1).

The *length* of a finite run $w = s_0, a_1, \dots, s_k$ is the number $\text{len}(w) = k$, i.e. the number of actions performed along w . The length of an infinite run is ∞ , and the lengths of (finite or infinite) histories are defined likewise. We denote by $\text{Run}_{\mathcal{M}}$ and $\text{FHist}_{\mathcal{M}}$ the sets of all runs and finite histories in \mathcal{M} , respectively.

Policies A *policy* (or a *policy*) in POMDP \mathcal{M} is a function σ of type $\text{FHist}_{\mathcal{M}} \rightarrow \mathcal{D}(A)$. Intuitively, policies are abstractions of controllers for the system modelled by \mathcal{M} : the control is exerted by choosing a suitable action in every decision step, depending on the history of the system's evolution. A run $w = s_0, a_1, s_1, \dots$ *conforms* to a policy σ if for all $0 \leq i < \text{len}(w)$ the distribution $\sigma(O(s_0, a_1, \dots, s_i))$ assigns positive probability to action a_{i+1} .

Semantics of POMDPs The behaviour of \mathcal{M} under a policy σ can be intuitively described as follows: first, an initial state s_0 is sampled according to the initial distribution λ . Then the system evolves in discrete steps. In a step $i \geq 0$, let $w_i = s_0, a_1, s_1, a_2, \dots, a_i, s_i$ be the current finite run, i.e. the sequence of traversed states and chosen actions up to the i -th step (we have $w_0 = s_0$). An action a_i is sampled according to the distribution $\sigma(O(w_i))$, and then a successor state s_{i+1} is sampled according to the distribution $\delta(s_i, a_i)$. In the next step the same procedure is performed with run $w_{i+1} = s_0, a_1, s_1, a_2, \dots, a_{i+1}, s_{i+1}$, etc. The process evolves in this manner *ad infinitum*. This intuitive description can be formalized by constructing a suitable probability measure \mathbb{P}^σ assigning probabilities to sets of infinite runs in \mathcal{M} . The construction of \mathbb{P}^σ is standard [6]. We denote by \mathbb{E}^σ the expected value operator induced by \mathbb{P}^σ .

Objectives An objective is a mathematical formalization of a desired behaviour of a system modeled by a POMDP. In this paper we study POMDPs that com-

bine *reachability*, *stochastic shortest path*, and *energy* objectives.

- A *reachability* objective is given by a set $T \subseteq S$ of *target* states. A run s_0, a_1, s_1, \dots satisfies such an objective if it eventually reaches a state from T , i.e. if $s_i \in T$ for some $i \geq 0$. We denote by $Reach_T$ the set of all infinite runs that satisfy a reachability objective with target set T .
- A *total cost* objective is given by a tuple (T, c) , where T is again a set of target states and $c: S \times A \rightarrow \mathbb{N}$ is a *cost function* assigning a positive integer cost to every state-action pair. Total cost is a quantitative objective, i.e. instead of saying that a run satisfies the objective or not, we measure the "quality" of a run by assigning a number to it. Here we assign to an infinite run $w = s_0, a_1, s_1, \dots$ its total cost $TC_T^c(w) = \sum_{i=1}^m c(s_{i-1}, a_i)$, where $m = \inf\{j \geq 0 \mid s_j \in T\}$. (We stipulate that an empty sum equals zero. Note that if $m = \infty$, then $TC_T^c(w) = \infty$.)
- An *energy* objective is given by a tuple (E, cap, T) , where $E: A \times Z \rightarrow \mathbb{Z}$ is a function assigning a *resource change* to every action-observation pair, $cap \in \mathbb{N}_0$ is a non-negative *capacity*, and T is a set of target states. For $(s, a, n) \in S \times A \times [cap]$ we define a *one-step resource update* $EnUp^{cap}(s, a, n) = \min\{cap, n + E(a, O(s))\}$. For a run $w = s_0, a_1, s_1, \dots$ we put an *energy level* after $i \in \mathbb{N}_0$ steps along w , where $0 \leq i \leq \text{len}(w)$, to be a number $EL_E^{cap}(w, i)$ defined inductively as follows: $EL_E^{cap}(w, 0) = cap$ and for $i \geq 1$ we put $EL_E^{cap}(w, i) = EnUp^{cap}(s_{i-1}, a_i, EL_E^{cap}(w, i-1))$. In other words, we assume that the resource level is initially at full capacity and is then changed by performing various actions. Should the resource level rise above cap , the excess amount is immediately discarded. An infinite run $w = s_0, a_1, s_1, \dots$ satisfies an energy objective given by (E, cap, T) if $EL_E^{cap}(w, i) > 0$ for all $0 < i \leq m$, where $m = \inf\{j \geq 0 \mid s_j \in T\}$. We denote the set of all such satisfying infinite runs by $Safe_{E,T}^{cap}$.

Remark 2. Note that per our definition the resource level at every step is perfectly observable. This is a reasonable assumption whenever the modelled energy-constrained agent is equipped with sufficiently precise charge/fuel sensors. In this our model resembles mixed-observability POMDPs [40, 2], and indeed in the next section we will present a transformation of POMDPs with energy constraints into standard POMDPs in which resource levels are a fully observable component of each state. However, mixed observability is used to enhance the performance of exact and point-based algorithms, while we aim for solution via simulation-based techniques, namely RTDP-Bel. Since, in the words of [2], online techniques cannot be probably adapted

to benefit from mixed observability, we stick to standard POMDP formulations. For further applications of mixed observability, see, e.g. [15, 13].

Computational Tasks Given a POMDP \mathcal{M} , a set of states T , an resource change function E , and a capacity cap , we define the set of energy-safe policies $\text{EnSafe}_T^{\mathcal{M}}(E, cap)$. A policy σ belongs to the set $\text{EnSafe}_T^{\mathcal{M}}(E, cap)$ if for all infinite runs w conforming to policy σ we have $w \in \text{Safe}_{E,T}^{cap}$. Given a policy $\sigma \in \text{EnSafe}_T^{\mathcal{M}}(E, cap)$ we define the value of σ as the expectation $\text{Val}(\sigma) = \mathbb{E}^\sigma[TC_T^c]$.

We are interested in minimizing the expected cost till the target set T is reached while keeping the energy level positive, i.e., we are interested in approximating the following value:

$$\text{optCost} = \inf_{\sigma \in \text{EnSafe}_T^{\mathcal{M}}(E, cap)} \text{Val}(\sigma)$$

We aim to solve the following computational problems:

1. The *qualitative energy-reachability problem* asks whether $\text{optCost} < \infty$.
2. The *quantitative energy-reachability problem* asks for a policy σ such that $\text{Val}(\sigma)$ approximates the value optCost .

Remark 3. We remark about POMDPs without energy constraints and the restrictions of the cost function:

1. The problem of approximating optimal cost optCost in POMDPs without energy constraints for positive costs was shown to be decidable in [17].
2. The problem of approximating optimal cost optCost for general costs (positive and negative) was proved to be undecidable in [17] already for POMDPs without energy constraints.

As policies are per definition infinite objects, in both the qualitative and quantitative problems we aim to compute their finite representations. One of the primary aims of this paper is to address the efficiency of such representations, our goal being to find succinct and/or human-readable ways to encode the computed policies.

4 Solving Energy-Reachability Problems

For the rest of the section let us fix a POMDP \mathcal{M} , target set T , functions c, E and a capacity cap . We will evaluate the complexity of presented algorithms in terms of $\|\mathcal{M}\|$ and $\|cap\|$. We assume that cap is represented in binary, i.e. cap is at most exponential in $\|cap\|$.

To solve both types of energy-reachability problems we construct a *product* POMDP \mathcal{M}_\times by encoding the

resource levels in \mathcal{M}_\times directly into the states. Formally, \mathcal{M}_\times has a set of states $S_\times = S \times [cap] \cup \{\perp\}$, where \perp is a newly added *sink* state, and the same set of actions as \mathcal{M} . A transition function δ_\times of \mathcal{M}_\times is defined as follows: for all $(s, n) \in S \times [cap]$, all $a \in A$ s.t. $EnUp^{cap}(s, a, n) \geq 1$, and all $\tilde{s} \in S_\times$ we have

$$\delta_\times(\tilde{s}[(s, n), a]) = \begin{cases} \delta(s'|s, a) & \text{if } \tilde{s} = (s', EnUp^{cap}(s, a, n)) \\ 0 & \text{otherwise;} \end{cases}$$

and for every other $\hat{s} \in S_\times$, $a \in A$ the distribution $\delta_\times(\hat{s}, a)$ is Dirac, assigning 1 to state \perp . The set of observations in \mathcal{M}_\times is $Z_\times = Z \cup \{\perp\}$, and observation function O_\times is such that $O_\times(s, n) = O(s)$ for each $(s, n) \in S \times [cap]$ and $O_\times(\perp) = \perp$. Finally, the initial distribution λ_\times assigns to each tuple of the form (s, cap) probability $\lambda(s)$, and 0 to all other states. We also extend the reachability and total cost objectives to \mathcal{M}_\times by defining a new target set $T_\times = T \times [cap]$ and cost function c_\times such that $c_\times((s, n), a) = c(s, a)$ and $c_\times(\perp, a) = 1$, for every action a .

It is straightforward to verify that POMDP \mathcal{M}_\times can be automatically constructed in time polynomial in $\|\mathcal{M}\|$ and exponential in $\|cap\|$.

There is a natural correspondence between runs, histories, and policies in \mathcal{M} and \mathcal{M}_\times which preserves the properties related to our objectives. In particular, for every policy $\sigma \in \text{EnSafe}_T^{\mathcal{M}}(E, cap)$ that almost surely reaches T one can construct a policy $\tilde{\sigma}$ in \mathcal{M}_\times such that $\tilde{\sigma}$ almost surely reaches T_\times and $\mathbb{E}^{\tilde{\sigma}}[TC_{T_\times}^{c_\times}] = \mathbb{E}^\sigma[TC_T^c]$; and vice versa, any policy in \mathcal{M}_\times that almost surely reaches target, can be transformed into a policy of the same expected cost in $\text{EnSafe}_T^{\mathcal{M}}(E, cap)$. If the policy σ to be transformed is finitely represented, the finite representation of the transformed policy $\tilde{\sigma}$ can be computed in time polynomial in $\|\mathcal{M}_\times\|$ and $\|\sigma\|$.

It follows that to solve the qualitative energy-reachability problem for \mathcal{M} it suffices to solve the qualitative reachability problem for \mathcal{M}_\times , i.e. compute a policy $\tilde{\sigma}$ such that $\mathbb{P}^{\tilde{\sigma}}(\text{Reach}_{T_\times}) = 1$. Algorithm solving the qualitative reachability problem based on *belief supports* was presented in [17, 3]. We briefly recall the approach: A belief support of a finite history $\rho = z_0, a_1, z_1, \dots, z_n$ is a set $B(\rho)$ of states in which the POMDP can be with positive probability after the sequence ρ is observed, i.e. $B(\rho) = \{s \in S \mid \exists w = s_0, a_1, s_1, \dots, s_n \in \text{Run}_{\mathcal{M}_\times} : \rho = O(w) \wedge s = s_n\}$. The algorithm computes for each $U \in BS(\mathcal{M}_\times) = \{U \subseteq S \mid \exists \rho \in FHist_{\mathcal{M}_\times} : B(\rho) = U\}$ a set of so-called *allowed* actions in U : intuitively, action a is allowed in U if playing action a in any situation where the observed finite history has belief support U results into situation in which the target set can still be reached with probability 1 by some policy. One can show that if there is a state $s \in \text{supp}(\lambda_0)$ such that $B(O(s))$ admits no allowed action, then no policy in \mathcal{M}_\times can reach the

target almost surely. Otherwise the algorithm outputs a policy σ_{all} which for each finite history ρ plays all actions allowed in $B(\rho)$ with uniform probability. It can be proved that for σ_{all} it holds $\mathbb{P}^{\sigma_{all}}(\text{Reach}_{T_\times}) = 1$.

The running time of the algorithm and the space needed to represent σ_{all} is dominated by a polynomial in the size of $BS(\mathcal{M}_\times)$, i.e. in the number of reachable belief supports. This number can be trivially bounded by $2^{|S_\times|}$, which is an expression doubly exponential in $\|cap\|$. However, from the construction of \mathcal{M}_\times we get the following improved bound:

Lemma 1. *It holds $|BS(\mathcal{M}_\times)| \leq 2^{|S| \cdot cap}$.*

As a consequence we get the following.

Theorem 1. *The qualitative energy-reachability problem for POMDPs is EXPTIME-complete.*

Proof. The upper bound follows from Lemma 1, and from the complexity of constructing \mathcal{M}_\times and translating its policies to \mathcal{M} . The lower bound follows from EXPTIME-hardness of qualitative reachability in POMDPs [19]. \square

To solve the quantitative energy-reachability problem, we again use an algorithm for POMDPs without energy, namely the one from [17], and apply it to \mathcal{M}_\times . The algorithm, which assumes that the sets of allowed actions were already computed via the aforementioned method, finds a policy of small cost that almost surely reaches T_\times , using a modified version of RTDP-Bel [8]. RTDP-Bel is an adaptation of the *real-time dynamic programming* value iteration [4] to POMDPs. It is an approximative method which does not guarantee convergence to optimum, but it is known to produce near-optimal policies on many instances where optimal costs can be computed using exact methods [17]. Hence, results produced by RTDP-Bel are a useful yardstick against which policies obtained by other methods can be compared. Due to the absence of guarantees we do not investigate the theoretical complexity of the algorithm. Its experimental evaluation can be found in Section 6.

The policy output by modified RTDP-Bel bases its decision in every step on the current *belief*, i.e. the probability distribution over the set of states representing the likelihood of being in particular states given the current history of states and observations [33]. As the space of beliefs is continuous, the policy operates on its discretized version, which allows it to be represented by a finite table storing one action per discretized belief. Not all beliefs have to be stored in the table: RTDP-Bel can converge to optimum without considering all reachable beliefs. Still, memory required to store the table might be too large for the policy to be understandable. In the next section we present a framework for converting table-represented policies into a more succinct and human-readable form.

5 Succinct Representation of Policies

A policy in \mathcal{M}_\times computed by the RTDP-Bel is a function which to every belief assigns an action to be taken. Formally, a belief is a probability distribution b on S_\times such that $\text{supp}(b) \subseteq O_\times^{-1}(z)$ for some observation z .

As indicated above, RTDP-Bel considers only discretized beliefs, that is beliefs whose probabilities are rounded to a finite mesh. For technical reasons, the RTDP-Bel represents such discretized beliefs as vectors of non-negative integers from an interval $[0, B]$ where B is a bound which determines the precision of the approximation.

Thus each belief in \mathcal{M}_\times can be represented as a vector $b \in \mathbb{Z}^{S+1}$ whose first $|S|$ components are integers from $[0, B]$, and whose last component is in $[cap]$. Given such a belief b , the true probability of being in a state s with the energy n is (approximately) equal to b^s/B , where b^s is the component corresponding to state s .

For simplicity, we assume that actions are named in such a way that $A = \{0, 1, \dots, k_A\}$.

5.1 Decision Trees

There are numerous possibilities of succinctly representing sets of vectors of numbers (and functions on such sets) in a human readable form. One of the most popular formalisms suitable for this purpose are decision trees (DT see [44, 38]). We use DTs to represent functions of beliefs in POMDPs. For convenience, we follow closely the definition of DT used in [10]. Let $V = \{v_1, \dots, v_d\}$ be a set of variable names.

Definition 1. A decision tree over the set of variables V is a tuple $\mathcal{T} = (Tr, \rho, \theta)$ where Tr is a finite rooted binary (ordered) tree with a set of inner nodes N and a set of leaves L , ρ assigns to every inner node a predicate of the form $[v_i \sim \text{const}]$ where $v_i \in V$, $\text{const} \in \mathbb{Z}$, $\sim \in \{\leq, <, \geq, >, =\}$, and θ assigns to every leaf a non-negative integer.

A DT \mathcal{T} over V determines a function $f : \mathbb{Z}^d \rightarrow \mathbb{Z}$ as follows: For a vector $\vec{v} = (\bar{v}_1, \dots, \bar{v}_n) \in \mathbb{Z}^d$, we find a path p from the root to a leaf ℓ such that for each inner node n on the path, the predicate $\rho(n)$ is satisfied by substitution $v_i = \bar{v}_i$ iff the first child of n is on p . Then we put $f(\vec{v}) = f(\bar{v}_1, \dots, \bar{v}_n) = \theta(\ell)$. In our setting the set of variable names is chosen so as to suitably characterize the current belief. Typically, one can put $V = S \cup \{\text{Energy}\}$, although different sets can be used as well. The domain of values assigned to leaves is the set of actions $A = \{0, 1, \dots, k_A\}$.

Training DT.

We describe the process of *learning a training set*, which can also be understood as storing the input/output behaviour of a function described by data. Assume that we are given a training sequence $\tau = (\bar{v}^1, f_1), \dots, (\bar{v}^k, f_k)$ (repetitions allowed!) that specifies the desired input/output behaviour, i.e. each $\bar{v}^i = (v_1^i, \dots, v_n^i) \in \mathbb{Z}^d$ is a training input and $f_i \in \mathbb{Z}$ is the expected output. The goal is to learn a DT which exhibits the input/output behaviour prescribed by the training sequence.

A standard process of learning according to the algorithm ID3 [44, 38] proceeds as follows:

1. Start with a single node (root), and assign to it the whole training sequence.
2. For a node n with a sequence $\tau = (\bar{v}^1, f_1), \dots, (\bar{v}^k, f_k)$,
 - (a) if all training examples in τ have the same expected output value (i.e. there is x such that $f_i = x$ for all i), set $\theta(n) = x$ and stop;
 - (b) otherwise,
 - choose a predicate with the “highest information gain” (with lowest entropy, see e.g. [38, Sections 3.4.1, 3.7.2]),
 - split τ (according to the inputs) into sequences satisfying and not satisfying the predicate, assign them to the first and the second child, respectively,
 - go to step 2 for each child.

Intuitively, the predicate with the highest information gain tends to make a clear cut among the classes.

In addition, the final tree can be *pruned*. This means that some leaves are merged, resulting in a smaller tree at the cost of some imprecision of storing. The pruning phase is quite sophisticated, hence for the sake of simplicity and brevity, we omit the details here. We use the standard C4.5 algorithm and refer to [45, 38]. In Section 6, we comment on effects of parameters used in pruning. We also use the CART algorithm [12] with so called Gini index instead of the information gain to select the best splits (there are also differences in pruning).

5.2 Learning a DT Policy

Our goal is to train a DT to succinctly represent a policy computed by RTDP-Bel. We use RTDP-Bel to generate a training set using the following procedure:

- Compute a policy σ using RTDP-Bel that solves the quantitative energy-reachability problem.
- Run a specified number m of simulations of σ , each of a fixed length ℓ . In every step of each simulation produce a new training instance $(b, \sigma(b))$ where
 - b is the current belief,

- $\sigma(b)$ is the action chosen by σ in the current step.

The above procedure generates a training sequence of pairs $\tau = (b_1, \sigma(b_1)), \dots, (b_k, \sigma(b_k))$ where $k = m \cdot \ell$. We feed this sequence into a learning algorithm for DT and obtain a decision tree approximating behaviour of σ on beliefs visited by the simulations.

Producing τ using RTDP-Bel should serve as a tool for detecting the most important decisions of σ . The intuition is that whenever a decision in a belief is made repeatedly in many simulations, it is worth remembering in the DT.

Execution of a DT Policy. The agent maintains the current belief over the state-space. The decision tree \mathcal{T} represents a function from beliefs to recommended actions as described in Section 5.1. In every step, the agent computes the value of the function for the current belief. Whenever a non-allowed action is recommended by the decision tree, we detect this and play all allowed actions uniformly at random, taking advantage of the precomputed set of allowed actions. Note that similar situation may also arise when executing a policy computed by RTDP-Bel: such a policy does not typically store decisions for all beliefs (see Section 4) and hence it might happen during its execution that no entry for the current belief can be found. Whatever action the implementation chooses in such a situation, the action must be allowed in the current belief-support, because after playing a non-allowed action, the set of target states would not be reached with probability 1. It is thus reasonable to assume that the precomputed information on allowed actions must be stored when executing any of the two types of policies. Hence, we view the difference between the size of the RTDP-Bel policy (i.e. the number of entries in its table) and the size of the DT learned from this policy (i.e. the number of its nodes) to be the primary measure of how much succinctness can be achieved by using decision trees.

Comparison with Finite-state Controllers. We illustrate the conceptual differences between finite-state controllers and decision trees on a toy POMDP example. Figure 1 depicts the *Energy-constrained Tiger POMDP* which is an extension of the famous Tiger problem introduced in [32]. Imagine an agent standing in front of two closed doors. Behind one of the doors is a tiger and behind the other is a treasure. If the agent opens the door with the tiger a huge cost is received. In all other situations no cost is received. After opening any of the doors the POMDP reaches a terminal configuration. The agent can also listen, in order to gain some information about the location of the tiger. Unfortunately, listening is not entirely accurate. There is a 15% chance that the agent will hear a tiger be-

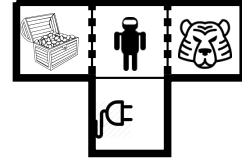


Figure 1: Energy-constrained Tiger Problem.

hind the left-hand door when the tiger is really behind the right-hand door, and vice versa. We consider two cases in the first one the listening action has no resource consumption and in the second case it has a resource consumption associated with the action. In the second case if the agent is running out of the resource, it may decide to recharge in a neighbouring location.

In the case without energy consumption, the optimal cost is 0, as for every $\varepsilon \geq 0$ there is a number T such that after performing L listening actions the robot builds enough confidence about the position of the tiger so as to choose the good door with probability at least $1 - \varepsilon$. At the same time, L listening actions are necessary to build such a confidence, and $L \rightarrow \infty$ as $\varepsilon \rightarrow 0$. To represent a policy that waits until enough confidence is built and then decides we can employ both FSCs and DTs. In the case of FSC the corresponding finite transducer needs to have number of states proportional to L , as it needs to count the number of listening actions so far as well as results of these actions. In a DT case the policy can be represented using a DT with 5 nodes, which is depicted in Figure 2. We use two variables *Tiger – Left* and *Tiger – Right*, that represent the probability that the tiger is behind the respective door. In each step the robot straightforwardly extracts these probabilities from its current belief and uses the decision tree to select an appropriate action. (A solid line to a successor is taken if the condition inside an inner node is satisfied, otherwise the dashed line is taken. The parameter δ represents the confidence level sufficient for door selection.) In particular, the size of the tree is independent of ε .

In the case with energy consumption, the FSC either needs to store the information on the current resource level in the transducer’s state, or, if the resource levels are contained in observations (as is the case in our reduction in the previous section), the FSC needs to have, in each of its states, at least one outgoing transition per each energy level (as there must be at least one outgoing transition for each possible observation). Thus, the size of such a controller is proportional also to the capacity *cap*. On the other hand, using a DT it suffices to slightly modify the tree in Figure 2. We need to use an additional variable representing the current resource level (which we assume is precisely known to the robot), and add a new root node which tests whether

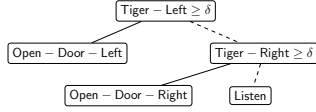


Figure 2: Decision Tree for the Tiger Problem.

the energy is greater than 1: if it is, the robot goes to a sub-tree depicted in Figure 2, otherwise it recharges in the neighbouring location.

6 Experimental results

We have extended the POMDP file format introduced in [14] with the constructs necessary to model resource consumption. We have implemented Algorithm 1 that implements the product construction of Section 4. We use a modified version of the RTDP-Bel POMDP solver [8] to solve these energy-constrained POMDPs and generate training data for machine learning tools (see the previous section).

Algorithm 1

Input: POMDP \mathcal{M} with energy-reachability objective

Output: A succinct policy σ if there exists one

$\mathcal{M}_\times \leftarrow \text{constructProduct}(\mathcal{M})$ ▷ Section 4

$\sigma_r \leftarrow \text{RTDP-Bel}(\mathcal{M}_\times)$

$\tau \leftarrow \text{trainingData}(\sigma_r)$ ▷ Section 5

$\sigma \leftarrow \text{trainTree}(\tau)$ ▷ Section 5

return: σ

Generating Training Data. In all the examples we consider we choose appropriate variables to characterize the beliefs, e.g. in grid-like environments we have variables for the current row, column, and current resource level (recall that resource level is perfectly observable). We parse the policy produced by RTDP-Bel to obtain a training sequence $(\vec{v}_1, a_1), (\vec{v}_2, a_2), \dots$ where each \vec{v}_i is a vector of variable values characterizing a belief to which the RTDP-Bel policy assigns action a_i .

Decision Tree Learning. We use three different methods to obtain decision trees for our examples:

1. In the first scenario decision trees are constructed using the Weka machine learning package [27]. The Weka suite offers various decision tree classifiers. We use the J48 classifier, which is an implementation of the C4.5 algorithm [45];
2. We use the package *rpart* [58] from R [46] which implements the CART algorithms of [12]. We use the Gini index as default for selecting the best splits. We have experimented with tree pruning using complexity parameters (see [58]).

3. Finally, we constructed trees using the package *tree* [47] of R which implements algorithms of [48]. In this case, the default measure for selecting splits is the deviance (see [48]). Note that the results of tree and rpart packages are usually very different.

We experimented on two well-known examples of POMDPs naturally extended with resource levels. Typically, we assume that all the actions of the agent decrease the resource level and there are specific recharging locations in the area that restore the resource level to the maximum capacity. The POMDP examples we considered are the following: (A) We consider the Hallway example from [36, 56, 53, 8]. (B) We consider the RockSample example from [8, 53].

6.1 Succinct Policies: Example

In this part we discuss an example of a succinct human readable policy for the Hallway examples

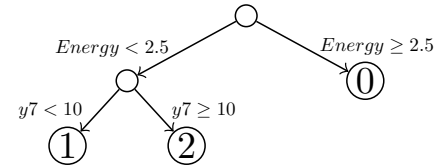


Figure 3: policy for Hallway on an 8x8 grid with $cap = 10$.

Example 1. Figure 3 shows a decision tree computed via *Tree* package for an instance of a Hallway example, which models a robot navigation in a maze. We use variable names $x_1, \dots, x_8, y_1, \dots, y_8, Energy$, where the values x_n, y_n represent the probability that the x or y coordinate, respectively, of the robot is equal to n (we have $B = 20$, i.e. value 20 represents probability 1). For better readability, predicates are not contained within nodes, they label edges instead. To execute a policy represented by the tree, the robot looks, in every step, on its current belief. If the current resource level is at least 3 (recall that resource levels are integers), it performs action 0 ("move forward"). Otherwise the robot checks the probability of the current y coordinate being 7. If it is smaller than $\frac{1}{2}$, the robot turns left (action 1), otherwise it turns right (action 2). If at any point the action recommended by the tree is not allowed, the robot chooses an allowed action uniformly at random.

6.2 Discussion on Experimental results.

We present the results of our approach in Table 1. Every entry contains the following information: (i) the name of the benchmark; (ii) the size of the state space; (iii) the maximum resource level cap ; (iv) the size of the the product state space S_\times after a preprocessing

Name	$ S $	cap	$ S_{\times} $	$Val(\sigma_{all})$	RTDP-Bel		Weka		RPart		Tree	
					Mem. size	Val	DT. size	Val	DT. size	Val	DT. size	Val
Hallway5x5	51	10	197	<i>568.001</i>	216	10.998	–	–	–	–	–	–
Hallway6x6	83	10	443	<i>862.746</i>	314	15.143	1	32.390	17	22.143	15	31.596
Hallway8x8	155	10	961	<i>954.160</i>	537	21.020	5	44.399	21	26.396	5	52.689
Hallway10x10	259	15	2437	<i>984.647</i>	2441	23.027	–	–	25	60.643	55	<i>573.983</i>
Hallway13x13	435	10	2852	<i>958.284</i>	1684	31.492	–	–	–	–	51	58.395
Hallway17x17	835	10	5163	<i>797.728</i>	2777	26.904	15	46.627	7	41.820	9	40.480
Hallway17x17	835	15	8882	<i>716.796</i>	6305	27.159	9	59.022	7	51.880	51	38.400
Hallway19x19	1103	10	8891	<i>883.223</i>	6801	30.131	15	148.184	5	142.537	3	44.614
RockSample[3,4]	435	7	2403	12.891	1257	6.146	31	8.52	11	12.948	1	13.638
RockSample[3,5]	1011	7	5425	12.412	395	6.920	19	8.275	13	12.405	1	8.275
RockSample[4,4]	771	7	4297	14.852	570	9.408	25	14.285	11	14.964	1	15.644
RockSample[5,4]	1803	6	5679	16.290	421	9.578	25	17.192	11	17.044	1	17.192

Table 1: Experimental results

step which removes unreachable states; (v) the value of the policy σ_{all} that plays all allowed actions uniformly at random; (vi) RTDP-Bel entries, that present the size of the computed policy and its corresponding value; (vii) for Weka, RPart, Tree we present the size of the computed decision tree and the value of the corresponding policy. The entries labelled with “-” did not have a run that reaches the set of target states T , entries in italics do not reach the set of target states T with all the runs within the run cut-off length of 1000, i.e., the expected cost for the policies may be higher. The bold-faced entries present the best result among the three considered DT-learning tools.

The entries show that removing the unreachable states from the product POMDP \mathcal{M}_{\times} is efficient and allows scaling to larger instances, e.g., a naive product construction in entry Hallway 10x10 would yield a product POMDP with 3885 states compared to the 2437 reachable states.

In most of the cases our approach succeeded and computed a significantly smaller policy than the standard policy computed by RTDP-Bel. The computed succinct policies usually perform slightly worse than the optimal explicit policy, however still overwhelmingly outperform the naive policy σ_{all} that plays all actions uniformly at random. For instance, the tree for Hallway8x8 presented in Example 1 identifies two crucial decisions (one based on current resource level, and one on crossing a certain “latitude”) with which the policy already significantly outperforms σ_{all} and achieves performance relatively close to the RTDP-Bel policy of size 537. Interestingly, this suggests that POMDPs with SSP objectives exhibit a phenomenon known as *Pareto principle* [39], where a small fraction of decisions accounts for majority of optimization effort. Moreover, the learning techniques we used are typically able to identify such decisions. Even for RockSample examples, where there is not much room for imprecision (σ_{all} incurs only the double the cost incurred by the corresponding RTDP-Bel policy) the DT policies performed relatively well on

some instances.

Interestingly, no approach significantly outperforms any other and for each of the three considered approaches there are entries where it dominates. On a fraction of examples, such as Hallway 5x5 or on the large RockSample instances, the learned policies do not perform well. One way of improving the performance of DT policies would be to use more expressive variants of DTs, such as *linear DTs*, that can capture general linear dependencies between variables. Testing this conjecture would require extending the aforementioned DT-learning tools with the capability of learning linear DTs, which we deem to be a viable direction of future work.

7 Conclusion

In this work we have considered POMDPs with a set of target states, positive integer costs associated with every transition, and resource levels. We present a novel algorithm for solving POMDPs enhanced with resource levels based on the existing POMDP solvers and the RTDP method. We consider three different approaches to obtain succinct and human-readable policies. On two scalable domains from the existing literature we present succinct policies that perform only slightly worse than the optimal policy, while being significantly smaller.

References

- [1] Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall, 1999.
- [2] Mauricio Araya-López, Vincent Thomas, Olivier Buffet, and Francois Charpillet. A Closer Look at MOMDPs. In *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, volume 2, pages 197–204, Oct 2010.

- [3] Christel Baier, Marcus Grösser, and Nathalie Bertrand. Probabilistic ω -automata. *J. ACM*, 59(1):1:1—1:52, March 2012.
- [4] Andrew G Barto, Steven J Bradtke, and Satinder P Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72(1):81–138, 1995.
- [5] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995. Volumes I and II.
- [6] Patrick Billingsley. *Probability and Measure*. Wiley, 3rd edition, 1995.
- [7] Blai Bonet and Héctor Geffner. Learning Sorting and Decision Trees with POMDPs. In *ICML*, pages 73–81, 1998.
- [8] Blai Bonet and Héctor Geffner. Solving POMDPs: RTDP-Bel vs. Point-based Algorithms. In *IJCAI*, pages 1641–1646, 2009.
- [9] Craig Boutilier and David Poole. Computing optimal policies for partially observable decision processes using compact representations. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1168–1175, 1996.
- [10] Tomáš Brázdil, Krishnendu Chatterjee, Martin Chmelík, Andreas Fellner, and Jan Kreťinský. Counterexample Explanation by Learning Small Strategies in Markov Decision Processes. In *Proceedings of Computer Aided Verification - 27th International Conference (CAV 2015)*, volume 9206 of *LNCS*, pages 158–177. Springer, 2015.
- [11] Tomáš Brázdil, Krishnendu Chatterjee, Antonín Kučera, and Petr Novotný. Efficient Controller Synthesis for Consumption Games with Multiple Resource Types. In P Madhusudan and S A Seshia, editors, *Computer Aided Verification 2012*, volume 7358 of *LNCS*, pages 23–38, 2012.
- [12] Leo Breiman, Jerome H. Friedman, Charles J. Stone, and Richard A. Olshen. *Classification and Regression Trees*. Wadsworth, 1984.
- [13] Jesus Capitán, Luis Merino, and Anibal Ollero. Multi-Robot Coordinated Decision Making under Mixed Observability through Decentralized Data Fusion. In *Proceedings of the 11th International Conference on Mobile Robots and Competitions (Robotica 2011)*, 2011.
- [14] Anthony Cassandra. Pomdp-solve [Software, version 5.4]. <http://www.pomdp.org/>.
- [15] Iadine Chades, Josie Carwardine, Tara Martin, Samuel Nicol, Régis Sabbadin, and Olivier Buffet. MOMDPs: a solution for modelling adaptive management problems. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.
- [16] Arindam Chakrabarti and Luca De Alfaro. Resource interfaces. *Embedded Software*, pages 117–133, 2003.
- [17] Krishnendu Chatterjee, Martin Chmelík, Raghav Gupta, and Ayush Kanodia. Optimal Cost Almost-Sure Reachability in POMDPs. In *AAAI 2015*, 2015.
- [18] Krishnendu Chatterjee and Laurent Doyen. Energy and Mean-Payoff Parity Markov Decision Processes. In *Proceedings of MFCS 2011*, volume 6907 of *LNCS*, pages 206–218. Springer, 2011.
- [19] Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. Qualitative Analysis of Partially-Observable Markov Decision Processes. In *MFCS*, pages 258–269, 2010.
- [20] Krishnendu Chatterjee, Laurent Doyen, and Thomas A Henzinger. A survey of partial observation stochastic parity games. *Formal Methods in System Design*, 43(2):268–284, 2013.
- [21] Jerzy Filar and Koos Vrieze. *Competitive Markov Decision Processes*. Springer-Verlag, 1997.
- [22] Marek Grzes and Pascal Poupart. Incremental Policy Iteration with Guaranteed Escape from Local Optima in POMDP Planning. In Gerhard Weiss, Pinar Yolum, Rafael H. Bordini, and Edith Elkind, editors, *AAMAS*, pages 1249–1257. ACM, 2015.
- [23] Marek Grzes, Pascal Poupart, and Jesse Hoey. Controller Compilation and Compression for Resource Constrained Applications. In Patrice Perny, Marc Pirlot, and Alexis Tsoukias, editors, *ADT*, volume 8176 of *Lecture Notes in Computer Science*, pages 193–207. Springer, 2013.
- [24] Marek Grzes, Pascal Poupart, and Jesse Hoey. Isomorph-Free Branch and Bound Search for Finite State Controllers. In Francesca Rossi, editor, *IJCAI. IJCAI/AAAI*, 2013.
- [25] Ujjwal Das Gupta, Erik Talvitie, and Michael Bowling. Policy tree: Adaptive representation for policy gradient. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2547–2553, 2015.

- [26] Wolfgang A Halang. Simplicity considered fundamental to design for predictability. In *Design of Systems with Predictable Behaviour*, 2004.
- [27] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *ACM SIGKDD*, 11(1):10–18, 2009.
- [28] Eric A. Hansen. An Improved Policy Iteration Algorithm for Partially Observable MDPs. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *NIPS*, pages 1015–1021. The MIT Press, 1997.
- [29] Gerard J. Holzmann. The power of 10: rules for developing safety-critical code. *Computer*, 39(6):95–99, 2006.
- [30] Joshua D. Isom, Sean P. Meyn, and Richard D. Braatz. Piecewise Linear Dynamic Programming for Constrained POMDPs. In Dieter Fox and Carla P. Gomes, editors, *AAAI*, pages 291–296. AAAI Press, 2008.
- [31] Xiaofeng Jiang, Hongsheng Xi, Xiaodong Wang, and Falin Liu. Finding Optimal Observation-based Policies for Constrained POMDPs under the Expected Average Reward Criterion. *Automatic Control, IEEE Transactions on*, PP(99):1–1, 2015.
- [32] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.
- [33] Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.
- [34] Dongho Kim, Jaesong Lee, Kee-Eung Kim, and Pascal Poupart. Point-Based Value Iteration for Constrained POMDPs. In Toby Walsh, editor, *IJCAI*, pages 1968–1974. IJCAI/AAAI, 2011.
- [35] Hermann Kopetz. *Real-time systems: design principles for distributed embedded applications*. Springer Science & Business Media, 2011.
- [36] Michael L. Littman, Anthony R. Cassandra, and Leslie P. Kaelbling. Learning policies for partially observable environments: Scaling up. In *ICML*, 1995.
- [37] Nicolas Meuleau, Kee-Eung Kim, Leslie Pack Kaelbling, and Anthony R. Cassandra. Solving POMDPs by searching the space of finite policies. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 417–426. Morgan Kaufmann Publishers Inc., 1999.
- [38] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
- [39] Mark E.J. Newman. Power laws, Pareto distributions and Zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.
- [40] Sylvie C. W. Ong, Shao Wei Png, David Hsu, and Wee Sun Lee. Planning under Uncertainty for Robotic Tasks with Mixed Observability. *I. J. Robotic Res.*, 29(8):1053–1068, 2010.
- [41] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12:441–450, 1987.
- [42] Joelle Pineau, Geoff Gordon, Sebastian Thrun, and Others. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI*, volume 3, pages 1025–1032, 2003.
- [43] Martin L. Puterman. *Markov Decision Processes*. J. Wiley and Sons, 1994.
- [44] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- [45] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [46] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
- [47] Brian Ripley. *tree: Classification and Regression Trees*, 2015. R package version 1.0-36.
- [48] Brian D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.
- [49] Keith W. Ross and Bintong Chen. Optimal scheduling of interactive and noninteractive traffic in telecommunication systems. *Automatic Control, IEEE Transactions on*, 33(3):261–267, Mar 1988.
- [50] Keith W. Ross and Ravi Varadarajan. Markov decision processes with sample path constraints: the communicating case. *Operations Research*, 37(5):780–790, 1989.
- [51] Keith W. Ross and Ravi Varadarajan. Multichain Markov Decision Processes with a Sample Path Constraint: A Decomposition Approach. *Math. Oper. Res.*, 16(1):195–207, February 1991.

- [52] Richard D Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.
- [53] Trey Smith and Reid Simmons. Heuristic search value iteration for POMDPs. In *UAI*, pages 520–527. AUAI Press, 2004.
- [54] Edward J. Sondik. *The Optimal Control of Partially Observable Markov Processes*. Stanford University, 1971.
- [55] Edward J. Sondik. The Optimal Control of Partially Observable Markov Processes over the Infinite Horizon: Discounted Costs. *Operations Research*, 26(2):282–304, 1978.
- [56] Matthijs T.J. Spaan and Nikos Vlassis. A point-based POMDP algorithm for robot planning. In *ICRA*. IEEE, 2004.
- [57] Henry Spencer. The Ten Commandments for C Programmers (Annotated Edition). <https://www.lysator.liu.se/c/ten-commandments.html>. Accessed: November 16, 2016.
- [58] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2015. R package version 4.1-10.
- [59] Aditya Undurti and Jonathan P. How. An online algorithm for constrained POMDPs. In *ICRA*, pages 3966–3973. IEEE, 2010.
- [60] Aditya Undurti and Jonathan P. How. A decentralized approach to multi-agent planning in the presence of constraints and uncertainty. In *ICRA*, pages 2534–2539. IEEE, 2011.
- [61] Feng Wu, Nicholas R. Jennings, and Xiaoping Chen. Sample-Based Policy Iteration for Constrained DEC-POMDPs. In Luc De Raedt, Christian Bessière, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *ECAI*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 858–863. IOS Press, 2012.